UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE

TESI DI LAUREA

INFERENZA SU P(X<Y) BASATA SULLA VEROSIMIGLIANZA PROFILO MODIFICATA: IL MODELLO BI-NORMALE

RELATORE: Prof.ssa Laura Ventura

LAUREANDO: Luca Bogoni

ANNO ACCADEMICO 2010-2011

INDICE

INTRODUZIONE

Fin dall'antichità la guerra è stata uno dei più importanti motori per l'evoluzione di molti settori della scienza come la tecnologia, la logistica, la medicina e le scienze statistiche.

Nel corso della seconda guerra mondiale i sistemi radar consentivano di caratterizzare un oggetto secondo i parametri di latitudine, longitudine, altitudine e velocità. Tuttavia non erano sufficienti a classificarlo come amico, nemico o rumore. L'identificazione avveniva in questo modo: si scartavano gli oggetti definiti da parametri non compatibili (come velocità, quota) successivamente si passava ad analizzare gli oggetti ammissibili. In ambito aeronautico, l'operatore radar controllava se tra i piani di volo in suo possesso ce n'era uno compatibile con i parametri osservati. Se corrispondevano, allora veniva identificato e di conseguenza segnalato come amico, altrimenti la torre di controllo tentava di stabilire un contatto diretto.

E' naturale che tanto più tempestivo era il riconoscimento dei velivoli, tanto più si poteva evitare l'intervento degli intercettori, con un risparmio economico, oppure in caso di attacco si poteva ottenere un vantaggio strategico. Così durante la seconda guerra mondiale per la prima volta gli ingegneri adottarono un metodo grafico per valutare la bontà del criterio usato dagli operatori radar per stabilire il tipo di oggetto, soprattutto dopo l'attacco a Pearl Harbor.

Lo schema, che prese il nome di curva ROC, si può applicare a molti casi di classificazione binaria (si veda ad esempio Azzalini e Scarpa, 2004). La curva ROC si adatta a molti ambiti della scienza ed è semplice da costruire; infatti prevede di rappresentare la sensibilità (vero positivo) in ordinata e il complemento a uno della specificità (vero negativo) in ascissa. Uno dei metodi di sintesi più utilizzato derivato dalla curva ROC è l'area sottesa ad essa (AUC, *Area Under the ROC curve*), che consente di stimare la probabilità di assegnare un'unità statistica al suo gruppo reale di appartenenza, e di conseguenza valutare la bontà del metodo usato per la classificazione.

Un metodo parametrico classico per stimare l'AUC è basato sulla teoria della verosimiglianza (si veda ad esempio Pace e Salvan, 2001) applicata al modello sollecitazione-resistenza (*stress-strength model*). Questo metodo consente di valutare l'affidabilità di un componente considerando un test fisico in cui una variabile X rappresenta la sollecitazione e una variabile Y la resistenza che il componente dimostra. Se la sollecitazione supera la resistenza, cioè se $X>Y$, il componente si rompe: l'affidabilità si può quindi esprimere come la probabilità che non si rompa, cioè $P(X<Y)$.

Lo scopo di questa tesi è discutere l'utilizzo della verosimiglianza profilo modificata, un miglioramento della verosimiglianza profilo tramite un opportuno fattore di modificazione per ottenere un'inferenza più accurata sul parametro di interesse che in questa tesi è rappresentato da $P(X<Y)$.

CAPITOLO 1

Teoria della verosimiglianza

In questo capitolo richiamiamo brevemente alcuni concetti basilari della teoria della verosimiglianza, che torneranno utili per l'inferenza sulla quantità $P(X < Y)$. I principali riferimenti bibliografici per gli argomenti richiamati si trovano nel libro di Pace e Salvan (2001) e in Azzalini (2001).

1.1 Verosimiglianza

Consideriamo un insieme di dati osservati, il **campione,** che indicheremo con $y = (y_1, \cdots, y_n)$. L'assunto fondamentale su cui poggia l'inferenza statistica è che y costituisce una determinazione di una variabile casuale Y , e che si desidera utilizzarlo per trarre conclusioni sulla distribuzione $F(y; \theta)$ di Y , con $\theta \in \Theta \subseteq IR^p$, $p \geq 1$. Sia $p(y; \theta)$ la funzione di densità corrispondente. Quando le osservazioni sono indipendenti e identicamente distribuite, la funzione

$$L(\theta) = \prod_{i=1}^{n} p(y_i; \theta)$$

è detta **funzione di verosimiglianza** di θ basata sui dati y. Due funzioni di verosimiglianza che differiscono per una costante moltiplicativa si dicono **equivalenti**.

Spesso le procedure di inferenza basate su $L(\theta)$ sono espresse tramite la funzione di **log-verosimiglianza**, una trasformazione monotona crescente che rende la descrizione dei dati più semplice. La log-verosimiglianza è

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log\ p(y_i; \theta).$$

Due funzioni di log-verosimiglianza che differiscono per una costante additiva si dicono equivalenti. Un valore $\hat{\theta} \in \Theta$ tale che $L(\hat{\theta}) \geq L(\theta)$ per ogni $\theta \in \Theta$ è detto **stima di massima verosimiglianza**. In generale, non è detto che $\hat{\theta}$ esista o che sia unico. Nei modelli con verosimiglianza regolare, la stima di massima verosimiglianza si individua ponendo la derivata parziale prima della $l(\theta)$ uguale a zero, cioè risolvendo l'equazione di verosimiglianza

$$l_*(\theta) = 0,$$

con $l_* = \frac{\partial l(\theta)}{\partial \theta}$.

La **matrice di informazione osservata** è data dalle derivate parziali seconde di $l(\theta)$ cambiate di segno, ossia

$$j(\theta) = -l_{**}(\theta) = -\frac{\partial l_*(\theta)}{\partial \theta \partial \theta^T}.$$

Nei modelli statistici parametrici regolari lo stimatore di massima verosimiglianza $\hat{\theta}$ è consistente e la sua distribuzione asintotica è

$$\hat{\theta} \dot{\sim} N_p(\theta, j(\hat{\theta})^{-1}). \quad (1)$$

A partire dalla (1), si può ottenere la statistica test alla Wald

$$W_e(\theta) = (\hat{\theta} - \theta)\, j(\hat{\theta})(\hat{\theta} - \theta), \quad (2)$$

la cui distribuzione asintotica nulla è χ^2_p. Questa permette di costruire regioni di confidenza con livello nominale $1-\alpha$, di forma

$$\{\theta \in \Theta : W_e(\theta) < \chi^2_{p;1-\alpha}\},$$

dove $\chi^2_{p;1-\alpha}$ è il quantile $(1-\alpha)$ della distribuzione χ^2_p.

Una statistica test asintoticamente equivalente a $W_e(\theta)$ è data dalla statistica test **log-rapporto di verosimiglianza**

$$W(\theta) = 2\{l(\hat{\theta}) - l(\theta)\}. \quad (3)$$

Questa statistica permette di costruire regioni di confidenza con livello nominale 1-α, di forma

$$\{\theta \in \Theta : W(\theta) \leq \chi^2_{p;1-\alpha}\}.$$

Quando il parametro θ è scalare, si può considerare la statistica test **radice con segno** di $W(\theta)$, data da

$$r(\theta) = sgn(\hat{\theta} - \theta)\sqrt{W(\theta)},$$

la cui distribuzione asintotica nulla è $N(0,1)$. In questo caso, un intervallo di confidenza con livello $1-\alpha$ è

$$\{\theta \in \Theta : -Z_{1-\alpha/2} < r(\theta) < Z_{1-\alpha/2}\},$$

con $Z_{1-\alpha/2}$ quantile $(1-\alpha/2)$ della distribuzione $N(0,1)$.

1.2 Verosimiglianza profilo

In molte applicazioni si può essere interessati a fare inferenza su una sola componente del parametro θ. Sia θ partizionato come $\theta = (\psi, \lambda)$, dove ψ è un **parametro scalare di interesse**, mentre λ è un **parametro $(p-1)$ dimensionale di disturbo.** Per l'inferenza su ψ un procedimento di ampia applicabilità prevede di sostituire il parametro di disturbo con una sua opportuna stima.

In particolare, la **verosimiglianza profilo** prevede di sostituire il parametro di disturbo λ con la sua stima di massima verosimiglianza vincolata al parametro di interesse ψ fissato. Si ottiene

$$L_p(\psi) = L(\psi, \hat{\lambda}_\psi),$$

dove $\hat{\lambda}_\psi$ è la stima di massima verosimiglianza di λ con ψ fissato.

La verosimiglianza profilo ha delle proprietà che la rendono simile a una verosimiglianza propria:

i. La stima di massima verosimiglianza profilo di ψ basata su $L_p(\psi)$ coincide con la stima di massima verosimiglianza di ψ,$\hat{\psi}$, basata su $L(\psi, \lambda)$.

ii. Il log-rapporto di verosimiglianza profilo coincide con il log-rapporto di verosimiglianza basato su $L(\psi, \lambda)$ considerato per l'inferenza su ψ con λ ignoto, ossia

$$W_p(\psi) = 2\{l_p(\hat{\psi}) - l_p(\psi)\} = 2[l(\hat{\psi}, \hat{\lambda}) - l(\psi, \hat{\lambda}_\psi)]$$

con $l_p(\psi) = log L_p(\psi)$. La statistica $W_p(\psi)$ ha distribuzione asintotica nulla χ^2_1, sotto opportune assunzioni di regolarità. Pertanto, un intervallo di confidenza di livello nominale $1-\alpha$ per ψ ha forma

$$\{\psi: W_p(\psi) < \chi^2_{1;1-\alpha}\}.$$

La versione unilaterale di $W_p(\psi)$ è data dalla radice con segno di $W_p(\psi)$, ossia

$$r_p(\psi) = sgn(\hat{\psi} - \psi)\sqrt{W_p(\psi)},$$

la cui distribuzione asintotica nulla è $N(0,1)$. Un intervallo di confidenza per il parametro di interesse ψ di livello $1-\alpha$ ha quindi forma

$$\{\psi: -z_{1-\alpha/2} \leq r_p(\psi) \leq z_{1-\alpha/2}\}.$$

iii. L'informazione osservata profilo è data da

$$j_p(\psi) = -\left(\frac{\partial^2 l_p(\psi)}{\partial \psi^2}\right) = -\left(\frac{\partial^2 l(\psi,\lambda_\psi)}{\partial \psi^2}\right) = -l_{\psi\psi}(\psi,\hat{\lambda}_\psi) + l_{\psi\lambda}(\psi,\hat{\lambda}_\psi)\left(l_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)\right)^{-1} l_{\lambda\psi}(\psi,\hat{\lambda}_\psi),$$

dove $l_{\psi\psi}(\cdot)$, $l_{\psi\lambda}(\cdot)$, $l_{\lambda\psi}(\cdot)$ e $l_{\lambda\lambda}(\cdot)$ indicano gli elementi della partizione in ψ e λ di $l_{**}(\psi,\lambda)$. Il test alla Wald per ψ può essere espresso come

$$Z_{e_p}(\psi) = (\hat{\psi} - \psi) j_p(\hat{\psi})^{1/2},$$

con distribuzione asintotica nulla $N(0,1)$.

Come già accennato, la verosimiglianza profilo $L_p(\psi)$ non è una verosimiglianza in senso proprio. In particolare, quando la dimensione di λ è elevata e/o la numerosità campionaria è bassa, l'inferenza basata su $L_p(\psi)$ può risultare inadeguata. Per questo motivo in letteratura sono state proposte diverse versioni modificate della $L_p(\psi)$ (si veda ad esempio, Severini 2000).

1.3 Verosimiglianza profilo modificata

Le proprietà appena elencate rendono la verosimiglianza profilo interessante. Tuttavia, come già accennato, non si tratta di una verosimiglianza in senso proprio: infatti ricorrere a $L_p(\psi)$ equivale a comportarsi come se λ fosse noto e pari a $\hat{\lambda}_\psi$. Ciò può non essere appropriato se i dati sono carenti di informazione su λ, cosa che si verifica quando la dimensione di λ è elevata.

È conveniente quindi compensare tale mancanza con delle opportune modifiche della verosimiglianza profilo.

Una possibile modifica è stata presentata da Barndoff-Nielsen (1983,1988). La corrispondente verosimiglianza profilo modificata è definita come

$$L_{MP}(\psi) = L_p(\psi)M(\psi),$$

con $M(\psi)$ fattore di aggiustamento dato da

$$M(\psi) = \left|\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}}\right|^{-1} \left|j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)\right|^{-1/2}.$$

Il fattore di modificazione $M(\psi)$ è di ordine $O_p(1)$, per cui $L_p(\psi)$ e $L_{MP}(\psi)$ sono asintoticamente equivalenti al primo ordine. Inoltre è importante sottolineare che le statistiche (2) o (3) definite a partire dalla verosimiglianza profilo modificata, hanno la medesima distribuzione asintotica di $W_p(\psi)$ e di $Z_{e_p}(\psi)$. Tuttavia ci si possono attendere delle valutazioni più accurate, in particolare se $(p-1)$ è elevato (si veda, ad esempio, Pace e Salvan, 1996).

Un'altra versione della verosimiglianza profilo modificata può essere definita con il seguente fattore di aggiustamento (Ventura e Racugno, 2011)

$$M^*(\psi) = M(\psi)\pi(\psi) = \left|\frac{\partial \hat{\lambda}_\psi}{\partial \hat{\lambda}}\right|^{-1} \left|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\right|^{-1/2} \left|i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)\right|^{1/2},$$

dove $i_{\psi\psi.\lambda}(\psi, \lambda) = i_{\psi\psi}(\psi, \lambda) - i_{\psi\lambda}(\psi, \lambda) i_{\lambda\lambda}(\psi, \lambda)^{-1} i_{\lambda\psi}(\psi, \lambda)$, e $i_{\psi\psi}(\theta)$, $i_{\psi\lambda}(\theta)$ e $i_{\lambda\psi}(\theta)$ sono i blocchi della matrice di informazione attesa $i(\theta) = E(j(\theta))$.

CAPITOLO 2

La curva ROC

In tutti i campi della scienza vengono sistematicamente messe a punto e utilizzate procedure più o meno complesse e della più svariata natura, ma sempre ben codificate, allo scopo di verificare un'ipotesi. Tali procedure sono comunemente dette "test". I test possono essere classificati in due tipologie: "qualitativi", i quali restituiscono una risposta dicotomica (es. positivo/negativo, vero/falso) e "quantitativi" che producono risultati continui o discreti. La performance diagnostica di un test può essere valutata tramite un'analisi statistica basata sulla curva ROC (*Receiver Operating Characteristic*).

2.1 Classsificazione binaria

Il problema di base che genera incertezza nell'interpretazione di un test risiede nel fatto che, nella maggioranza dei casi, esiste una zona di sovrapposizione fra le distribuzioni dei risultati del test medesimo applicato in due popolazioni di persone, rispettivamente sani e malati.

Infatti, se le due popolazioni restituissero valori separati (Figura 1), allora sarebbe facile individuare sull'asse delle ascisse il valore di *cut-off* capace di discriminare con precisione assoluta le due popolazioni.

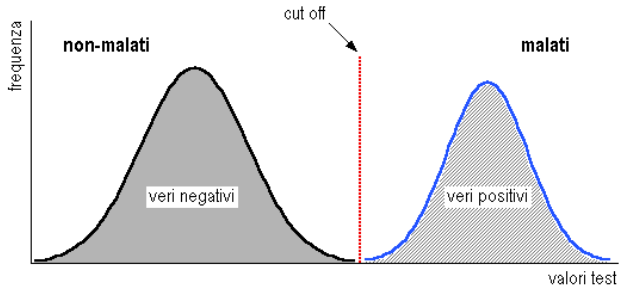

Figura 1: Distribuzione degli esiti di un ipotetico test nelle classi di individui malati e non malati, senza sovrapposizione inter-classe.

Nella pratica invece si verifica sempre una sovrapposizione più o meno ampia delle due distribuzioni (Figura 2) ed è perciò impossibile individuare sull'asse delle ascisse un valore di *cut-off* che consenta una classificazione perfetta, in modo da eliminare sia i falsi positivi che i falsi negativi.

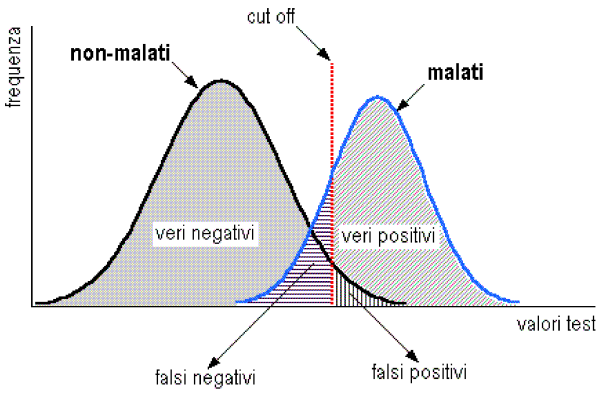

Figura 2: Distribuzione degli esiti di un ipotetico test nelle classi di individui malati e non malati, con sovrapposizione inter-classe.

La capacità diagnostica di un test ad un determinato valore di *cut-off* rappresenta la capacità di condurre ad una diagnosi positiva nei soggetti affetti da una determinata malattia e ad una diagnosi negativa nei soggetti non ammalati. Essa può essere valutata attraverso una semplice **tabella di contingenza** confrontando l'*output* del test in esame con il vero stato dei pazienti, in cui indichiamo con POSITIVO i pazienti affetti da malattia e con NEGATIVO i pazienti sani (vedi Tabella 1).

RISULTATO DEL TEST/ TIPO DI PAZIENTE	POSITIVO	NEGATIVO	
POSITIVO	TP(Veri Positivi)	FP(Falsi Positivi)	TP+FP
NEGATIVO	FN(Falsi Negativi)	TN(Veri Negativi)	FN+TN
	TP+FN	FP+TN	

Tabella 1: Tabella di contingenza

Nel caso binario i risultati di classificazione sono quattro: un'unità appartenente ai positivi è stata correttamente classificata come positiva, oppure un'unità è stata classificata come positiva mentre in realtà era negativa; ancora, un'unità appartenente ai negativi è stata correttamente classificata come negativa, oppure un'unità è stata classificata come negativa mentre in realtà era positiva.
Il confronto fra i risultati del test in esame e l'autentico stato di ogni individuo consente di stimare due importanti parametri: la **sensibilità** (Se), cioè la probabilità che un paziente malato risulti positivo, e la **specificità,** ossia la probabilità che un paziente sano risulti negativo:

$$Se = \frac{TP}{TP+FN} \quad \text{e} \quad Sp = \frac{TN}{FP+TN}.$$

Un test diagnostico risulta specifico al 100% quando tutti i sani risultano negativi.
Da notare che se un test presenta un'ottima specificità, allora è basso il rischio di Falsi Positivi, cioè di pazienti che pur presentando valori anomali non sono affetti dalla patologia che si sta ricercando; allo stesso modo se un test ha un'ottima sensibilità allora è basso il rischio di Falsi Negativi, cioè di pazienti che pur presentando valori anomali sono comunque affetti dalla patologia o dalla condizione che si sta ricercando.

2.2 Curva ROC

Uno strumento utilizzato per valutare l'adeguatezza di un test diagnostico, basato sulla sensibilità e specificità, è fornito dalla curva ROC.
Questa è stata introdotta nella II Guerra Mondiale nel contesto della teoria delle telecomunicazioni, in particolari segnali radio e poi è stata estesa in altri ambiti, particolarmente in controllo della qualità e in statistica medica.
Una considerazione che possiamo fare sulla Se e Sp è che sono fra loro inversamente proporzionali in rapporto alla scelta del valore di *cut-off*. Infatti possiamo ottenere uno dei seguenti effetti: aumento della Se e diminuzione della Sp, oppure diminuzione della Se e aumento della Sp. La "soglia discriminante ottimale", cioè il valore di *cut-off* che minimizza gli errori di classificazione, è pari al valore in ascissa corrispondente al punto di intersezione delle due distribuzioni.
Un metodo comune utilizzato per scegliere il valore di *cut-off* consiste nel fissare a priori il valore desiderato di specificità (generalmente >0.9) e quindi calcolare la corrispondente sensibilità del test. Questo metodo può generare due effetti collaterali negativi: il primo è rappresentato dal fatto che il test possa produrre risultati migliori tramite l'utilizzo di un valore di *cut-off* diverso da quello scelto. Il secondo legato al fatto che non si possa fare un raffronto in termini di performance tra due o più test valutati in base ad un solo valore di *cut-off*.

Queste osservazioni portano a due implicazioni:

1. È possibile scegliere un valore di *cut-off* che corrisponda ad un determinato valore di Se o Sp, ma non è detto che tale valore sia ottimale per gli scopi di studio;

2. La Se e la Sp associate ad un singolo valore di *cut-off* non rappresentano indicatori esaurienti della performance del test che si potrebbe ottenere con l'utilizzo di altri valori di *cut-off*.

L'analisi ROC si basa sullo studio della funzione che lega la probabilità di ottenere un risultato vero-positivo nella classe dei malati-veri (sensibilità) alla probabilità di ottenere un risultato falso-positivo nella classe dei non-malati (1-specificità). Si tratta di studiare i rapporti tra allarmi veri e allarmi falsi.

Questa relazione può essere raffigurata tramite una linea che si ottiene riportando, in un sistema di assi cartesiani e per ogni valore possibile di *cut-off*, la proporzione di veri positivi in ordinata e la proporzione di falsi positivi in ascissa. L'unione dei punti ottenuti riportando nel piano cartesiano ciascuna coppia (Se) e (1-Sp) genera una curva spezzata con andamento a scaletta (ROC *plot*). Per interpolazione, è possibile eliminare la scalettatura (*smoothing*) ed ottenere una curva (ROC *curve*) che rappresenta una stima basata sui parametri del data set sperimentale (Figura 3). Un buon test diagnostico avrà curva ROC il più possibile sopra la bisettrice.

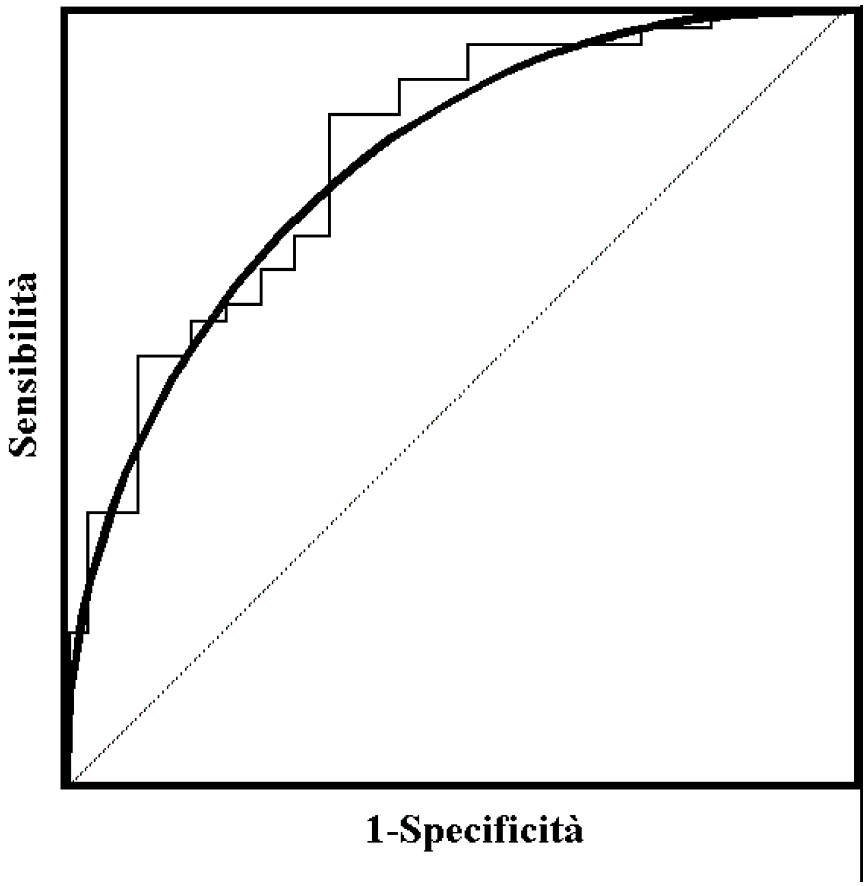

Fig. 3. Curva ROC prima e dopo interpolazione ("smoothing")

2.3 L’area sottesa alla curva ROC

Per avere una misura sintetica ed oggettiva dell’efficacia di un test si ricorre ad un indicatore di sintesi della curva ROC. Il più utilizzato è costituito dall’area sottesa dalla curva ROC, indicata con AUC (*Area Under the Roc Curve*).

Sia X la variabile che rappresenta la misura nel gruppo dei pazienti sani e Y quella dei pazienti malati. Si può esprimere l’AUC come

$$AUC = P(X < Y),$$

nota anche in letteratura come modello sollecitazione-resistenza (*stress-strength model*); si veda Kotz *et al*.(2003) .

Considerando la Figura 3, nel caso di un test perfetto, cioè non restituisce nessun falso positivo e nessun falso negativo (capacità discriminate = 100%), la AUC passa attraverso le coordinate $\{0; 1\}$ e il suo valore coincide con l’area dell’intero quadrato delimitato dai punti di coordinate $(0,0), (0,1), (1,0), (1,1)$, che assume valore 1 corrispondente alla probabilità del 100% di una corretta classificazione. La curva ROC priva di informazione è rappresentata dalla diagonale che passa per l’origine.

Ci sono due segmenti che hanno scarsa importanza per la valutazione dell’attitudine discriminante del test in esame e sono rappresentati dalle frazioni di curva che coincidono con l’asse delle ascisse e delle ordinate. Infatti, i corrispondenti valori possono essere scartati in quanto esistono altri valori di *cut-off* che forniscono una migliore Sp senza perdita di Se o viceversa una migliore Se senza perdita di Sp.

Infine bisogna dire che la valutazione di un test viene effettuata attribuendo la stessa importanza alla Se e alla Sp, mentre nei casi pratici si assegna un peso diverso ai parametri.

Per le variabili X e Y si possono fare assunzioni parametriche e non parametriche. In questa tesi si assume che X e Y sono variabili casuali indipendenti e appartenenti alla stessa famiglia di distribuzioni; in particolare si vedrà nello specifico il caso di due distribuzioni normali.

In generale si assume che X e Y siano due variabili casuali con funzione di densità, rispettivamente, $p_x(x;\theta_x)$ e $p_y(y;\theta_y)$, con $\theta_x \in \Theta_x \subseteq IR^{p_x}$ e $\theta_y \in \Theta_y \subseteq IR^{p_y}$.

L'AUC può essere espressa come una funzione del parametro $\theta = (\theta_x, \theta_y)$, tramite la relazione

$$AUC = AUC(\theta) = P(X < Y) = \int_{-\infty}^{+\infty} F_x(t;\theta_x) dF_y(t;\theta_y), \quad (4)$$

dove $F_x(\cdot)$ e $F_y(\cdot)$ sono le funzioni di ripartizione di X e Y, rispettivamente.

Espressioni teoriche per l'AUC sono disponibili con riferimento a diverse assunzioni distributive su X e Y(si veda Kotz. *et al*., 2003) .

2.4 Teoria della verosimiglianza

Siano $x = (x_1, \dots, x_{n_x})$ e $y = (y_1, \dots, y_{n_y})$ due campioni casuali semplici indipendenti di numerosità, rispettivamente n_x e n_y, tratti da X e Y. Allora una stima parametrica dell'AUC può essere ottenuta utilizzando la proprietà di equivarianza degli stimatori di massima verosimiglianza (vedi Pace e Salvan, 2001). Sia $\hat{\theta}$ la stima di massima verosimiglianza di θ, ottenuta massimizzando la funzione di verosimiglianza

$$L(\theta) = L(\theta_x, \theta_y) = \prod_{i=1}^{n_x} p_x(x_i;\theta_x) \prod_{j=1}^{n_y} p_y(y_i;\theta_y).$$

Allora la stima di massima verosimiglianza dell'AUC è

$$AUC = AUC(\hat{\theta}) = AUC(\hat{\theta}_x, \hat{\theta}_y).$$

Se poniamo $\theta = (\psi, \lambda)$, con $\psi = P(X < Y)$ e λ opportuno parametro di disturbo, si possono ottenere intervalli di confidenza per ψ a partire dalla verosimiglianza profilo, come visto nel Paragrafo 1.2. Ad esempio, un intervallo di confidenza di livello nominale $1-\alpha$ per l'AUC può essere costruito come

$$\{\psi \in (0,1): -Z_{1-\alpha/2} < r_p(\psi) < Z_{1-\alpha/2}\}.$$

Oppure utilizzando il test alla Wald otteniamo

$$\{\psi \in (0,1): -Z_{1-\alpha/2} < Z_{e_p}(\psi) < Z_{1-\alpha/2}\}.$$

Risultati più accurati si possono ottenere con le verosimiglianze profilo modificate

$$L_{MP}(\psi) = L_p(\psi)M(\psi)$$

e

$$L_{MP}{}^{*}(\psi) = L_p(\psi)M^{*}(\psi),$$

introdotte nel Paragrafo 1.3. In particolare, a partire dalle $L_{MP}(\psi)$ e $L_{MP}{}^{*}(\psi)$ si possono ottenere le statistiche $r_{MP}(\psi)$ e $r_{MP}{}^{*}(\psi)$ per costruire intervalli di confidenza approssimati, in modo analogo alla verosimiglianza profilo. Inoltre, stime puntuali di ψ possono essere ottenuta massimizzando $L_{MP}(\psi)$ e $L_{MP}{}^{*}(\psi)$. Il miglioramento nell'inferenza su ψ, che si ottiene con le verosimiglianze profilo modificate, sarà oggetto di studio nelle simulazioni effettuate nel Capitolo 3.

CAPITOLO 3

Il caso bi-normale

In questo capitolo si discutono degli studi di simulazione per valutare il miglioramento nell'inferenza sull'AUC introdotto dalle verosimiglianze profilo modificate. In particolare, in questo capitolo consideriamo X e Y distribuite secondo due distribuzioni normali, indipendenti.

Nella prima parte del capitolo presentiamo i passaggi algebrici per ottenere le quantità necessarie per l'inferenza sul parametro che rappresenta l'area sotto la curva ROC; nella seconda parte si procederà con la presentazione dei risultati delle simulazioni, effettuate utilizzando il software statistico R.

3.1 Distribuzione bi-normale

Siano $y = (y_1, \dots, y_{n_y})$ e $x = (x_1, \dots, x_{n_x})$ due campioni casuali semplici estratti da due variabili casuali normali indipendenti, rispettivamente, $X \sim$N(μ_1, σ^2) e Y$\sim$N(μ_2, σ^2), con $\mu_1 \epsilon\ IR$, $\mu_2 \epsilon\ IR$ e, le varianze assegnate uguali $\sigma^2 > 0$.

La verosimiglianza per $\theta = (\mu_1, \mu_2, \sigma^2)$ è

$$L(\theta) = L(\mu_1, \mu_2, \sigma^2) = \frac{1}{\sqrt{(\sigma^2)^{n_x}(\sigma^2)^{n_y}}}\ exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_x}(x_i - \mu_1)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_y}(y_i - \mu_2)^2\right\},$$

e la log-verosimiglianza è

$$l(\theta) = l(\mu_1, \mu_2, \sigma^2) = -\frac{1}{2}(n_x + n_y)\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n_x}(x_i - \mu_1)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_y}(y_i - \mu_2)^2,$$

dalla quale possiamo determinare gli stimatori di massima verosimiglianza. Infatti, derivando la log-verosimiglianza otteniamo le componenti della funzione punteggio, date da

$$\begin{cases} \dfrac{\partial l(\mu_1,\mu_2,\sigma^2)}{\partial \mu_1} = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=0}^{n_x}(x_i-\mu_1) \\ \dfrac{\partial l(\mu_1,\mu_2,\sigma^2)}{\partial \mu_2} = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=0}^{n_y}(y_i-\mu_2) \\ \dfrac{\partial l(\mu_1,\mu_2,\sigma^2)}{\partial \sigma^2} = -\dfrac{(n_x+n_y)}{2\sigma^2} + \dfrac{1}{2\sigma^2}\displaystyle\sum_{i=0}^{n_x}(x_i-\mu_1)^2 + \dfrac{1}{2\sigma^2}\sum_{i=0}^{n_y}(y_i-\mu_2)^2, \end{cases}$$

le cui soluzioni sono:

$$\hat{\mu}_1 = \bar{x},$$

$$\hat{\mu}_2 = \bar{y},$$

$$\hat{\sigma}^2 = \frac{1}{n_x+n_y}\left[\sum_{i=1}^{n_x}(x_i-\hat{\mu}_1)^2 + \sum_{i=1}^{n_y}(y_i-\hat{\mu}_2)^2\right].$$

Definiamo i nuovi parametri λ_1, λ_2 e ψ in funzione di μ_1, μ_2 **e** σ^2, come $\psi = \Phi\left(\frac{\mu_1-\mu_2}{\sqrt{2\sigma^2}}\right)$, $\lambda_1 = \mu_2$ *e* $\lambda_2 = \sigma$. Definiano, inoltre, le numerosità campionarie delle due distribuzioni come $n = n_x$ e $m = n_y$.

La funzione di log-verosimiglianza per (ψ, λ) diventa:

$$l(\psi,\lambda) = -(n+m)\left[\log\left(\lambda_2 + \frac{\tilde{\lambda}_2^2}{2\lambda_2^2}\right)\right] - \frac{1}{2\lambda_2^2}\left[m(\hat{\lambda}_1-\lambda_1)^2 + n(\hat{\lambda}_2\tilde{\psi}+\hat{\lambda}_1-\lambda_2\psi-\lambda_1)^2\right].$$

Da questa funzione possiamo calcolare le derivate parziali prime che risultano:

$$\begin{cases} l_{\lambda_1} = \frac{1}{\lambda_2^2}[\lambda_1(n+m) + \hat{\lambda}_1(n+m) + n(\hat{\lambda}_2\hat{\psi} - \lambda_2\psi)] \\ l_{\lambda_2} = -\frac{(n+m)}{\lambda_2} + \frac{1}{\lambda_2^3}\left[(n+m)\hat{\lambda}_2^2 + m(\hat{\lambda}_1 - \lambda_1)^2 + n(\hat{\lambda}_2\hat{\psi} + \hat{\lambda}_1 - \lambda_2\psi - \lambda_1)^2\right] + \frac{1}{\lambda_2^2}[n(\hat{\lambda}_2\hat{\psi} + \hat{\lambda}_1 - \lambda_2\psi - \lambda_1)\psi] \\ l_{\psi} = \frac{n}{\lambda_2^2}(\hat{\lambda}_2\hat{\psi} + \hat{\lambda}_1 - \lambda_2\psi - \lambda_1) \end{cases}$$

Derivando la funzione punteggio si ottengono gli elementi necessari per il calcolo della matrice di informazione osservata, la quale è composta dalle derivate seconde e derivate seconde cambiate di segno:

$$l_{\lambda_1\lambda_1} = -\frac{(n+m)}{\lambda_2^2}$$

$$l_{\lambda_1\lambda_2} = l_{\lambda_2\lambda_1} = -\frac{2}{\lambda_2^3}\left[(n+m)\hat{\lambda}_2^2 + n(\hat{\lambda}_2\hat{\psi} + \hat{\lambda}_1 - \lambda_2\psi - \lambda_1)\right] + \frac{1}{\lambda_2^2}[-\psi m]$$

$$l_{\lambda_2\lambda_2} = \frac{(n+m)}{\lambda_2^2} - \frac{3}{\lambda_2^2}\left[(n+m)\hat{\lambda}_2 + m(\hat{\lambda}_1 - \lambda_1)^2 + n(\hat{\lambda}_2\hat{\psi} + \hat{\lambda}_1 - \lambda_2\psi - \lambda_1)^2\right] + \frac{1}{\lambda_2^2}[-n\psi^2] - \frac{4}{\lambda_2^3}[n\psi(\hat{\lambda}_2\hat{\psi} + \hat{\lambda}_1 - \lambda_2\psi - \lambda_1)]$$

$$l_{\psi\psi} = -n$$

$$l_{\psi\lambda_1} = l_{\lambda_1\psi} = -\frac{n}{\lambda_2}$$

$$l_{\psi\lambda_2} = l_{\lambda_2\psi} = -\frac{n}{\lambda_2^2}(\hat{\lambda}_2\hat{\psi} + \hat{\lambda}_1 - \lambda_2\psi - \lambda_1) - \frac{n\psi}{\lambda_2}.$$

La matrice di informazione è:

$$j(\lambda_1, \lambda_2, \psi) = -\begin{pmatrix} l_{\lambda_1\lambda_1} & l_{\lambda_1\lambda_2} & l_{\lambda_1\psi} \\ l_{\lambda_2\lambda_1} & l_{\lambda_2\lambda_2} & l_{\lambda_2\psi} \\ l_{\psi\lambda_1} & l_{\psi\lambda_2} & l_{\psi\psi} \end{pmatrix}.$$

Con tali elementi si ottiene $j_p(\psi)$ e quindi la statistica Z_{e_p}. In particolare, il test alla Wald profilo risulta dato da

$$Z_{e_p} = (\hat{\psi} - \psi) j_p(\hat{\psi})^{1/2}$$

$$\text{con } j_p(\hat{\psi}) = \frac{\hat{G}}{(n\hat{\lambda}_2)^2 + n_x n_y (\hat{\lambda}_1 \hat{\lambda}_2 - \hat{D}(\hat{\theta}))^2},$$

dove $\hat{G} = 2nm_x n_y \hat{\lambda}_2^2 \hat{d}^2$ e $\hat{D} = \frac{\partial \phi^{-1}(\hat{\psi})}{\partial \hat{\psi}}$.

La verosimiglianza profilo modificata di Barndoff-Nielsen è

$$L_{MP}(\psi) = L_p(\psi) \left| \frac{\partial \hat{\lambda}_\psi}{\partial \hat{\lambda}} \right| \left| j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) \right|^{1/2},$$

dove $j_{\lambda\lambda}$ è una matrice composta dalle derivate parziali seconde:

$$j_{\lambda\lambda} = - \begin{bmatrix} l_{\lambda_1\lambda_1} & l_{\lambda_1\lambda_2} \\ l_{\lambda_2\lambda_1} & l_{\lambda_2\lambda_2} \end{bmatrix}$$

L'altra versione della verosimiglianza profilo modificata è data invece da

$$L_{MP}{}^*(\psi) = L_{MP}(\psi) \left| i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi) \right|^{1/2},$$

dove $i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)$ è data da

$$i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi) = i_{\psi\psi}(\psi, \hat{\lambda}_\psi) - i_{\psi\lambda}(\psi, \hat{\lambda}_\psi) i_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)^{-1} i_{\lambda\psi}(\psi, \hat{\lambda}_\psi),$$

le cui componenti sono gli elementi dalle matrice di informazione attesa

$$i(\psi,\lambda)=\begin{bmatrix} i_{\psi\psi}(\psi,\lambda) & i_{\psi\lambda_1}(\psi,\lambda) & i_{\psi\lambda_2}(\psi,\lambda) \\ i_{\lambda_1\psi}(\psi,\lambda) & i_{\lambda_1\lambda_1}(\psi,\lambda) & i_{\lambda_1\lambda_2}(\psi,\lambda) \\ i_{\lambda_2\psi}(\psi,\lambda) & i_{\lambda_2\lambda_1}(\psi,\lambda) & i_{\lambda_2\lambda_2}(\psi,\lambda) \end{bmatrix},$$

con

$$i_{\psi\psi}=n$$

$$i_{\psi,\lambda_1}=i_{\lambda_1,\psi}=\frac{n}{\lambda_2}$$

$$i_{\psi,\lambda_2}=i_{\lambda_2,\psi}=\frac{n\psi}{\lambda_2}$$

$$i_{\lambda_1\lambda_1}=\frac{(n+m)}{\lambda_2^2}$$

$$i_{\lambda_1\lambda_2}=i_{\lambda_2\lambda_1}=\frac{\psi m}{\lambda_2^2}$$

$$i_{\lambda_2\lambda_2}=\frac{2(n+m)}{\lambda_2^2}+\frac{n\psi^2}{\lambda_2^2}$$

3.2 Studio di simulazione

Lo studio di simulazione presentato è composto da una prima fase in cui si considera una stima puntuale di ψ, nella seconda si trattano le stime intervallari.

Con questo studio si vuole studiare il comportamento delle stime di ψ basate sulle $L_p(\psi)$, $L_{mp}(\psi)$, $L_{mp}{}^*(\psi)$, variando sia la numerosità campionaria sia il vero valore di ψ. I risultati della simulazione sono riassunti nella Tabella 3.1, dove per ogni cella viene indicata la media e la deviazione standard, tra parentesi, dei vettori delle stime simulando 1000 volte due campioni di numerosità n_x e n_y, con ψ fissato a 0.8, 0.9 e 0.95.

Nella tabella vengono prese in considerazione le stime basate sui tre metodi: il primo valore basato sulla verosimiglianza profilo L_p, il secondo sulla verosimiglianza profilo modificata L_{mp} di Barndoff-Nielsen e il terzo basato sulla seconda versione della verosimiglianza profilo modificata L_{mp}^*.

(n_x, n_y)	*test*	$\psi = 0.8$	$\psi = 0.9$	$\psi = 0.95$
(5,5)	$L_p(\psi)$	0.844(0.917)	0.941(0.950)	0.977(0.998)
	$L_{mp}(\psi)$	0.837(0.858)	0.933(0.968)	0.971(0.990)
	$L_{mp}^*(\psi)$	0.822(0.813)	0.922(0.916)	0.963(0.986)
(10,10)	$L_p(\psi)$	0.819(0.575)	0.917(0.576)	0.952(0.624)
	$L_{mp}(\psi)$	0.815(0.561)	0.913(0.533)	0.963(0.646)
	$L_{mp}^*(\psi)$	0.809(0.546)	0.908(0.539)	0.959(0.631)
(20,20)	$L_p(\psi)$	0.813(0.340)	0.909(0.399)	0.958(0.426)
	$L_{mp}(\psi)$	0.811(0.337)	0.908(0.391)	0.954(0.411)
	$L_{mp}^*(\psi)$	0.807(0.332)	0.904(0.386)	0.957(0.415)

Tabella 3.1: Studio Monte Carlo per la stima puntuale

Si può vedere come lo stimatore ottenuto dalla verosimiglianza profilo modificata L_{mp}^* è affidabile in termini di media ma anche in termini di deviazione standard (Figura 3.1). Infatti dai valori in tabella si nota come siano quelli che si avvicinano maggiormente al valore fissato di ψ. Nella Figura 3.1 possiamo vedere la distribuzione degli stimatori ottenuti dalle tre versioni della verosimiglianza profilo, con ψ fissato a livello 0.80. Notiamo che la stima ottenuta con L_{mp}^* si avvicina di più al valore 0.80 rispetto alle altre due versioni.
I grafici successivi (Figura 3.2, Figura 3.3) hanno lo stesso significato; l'unico cambiamento è fatto sul parametro ψ, il quale viene posto uguale a 0.90 e 0.95, rispettivamente.

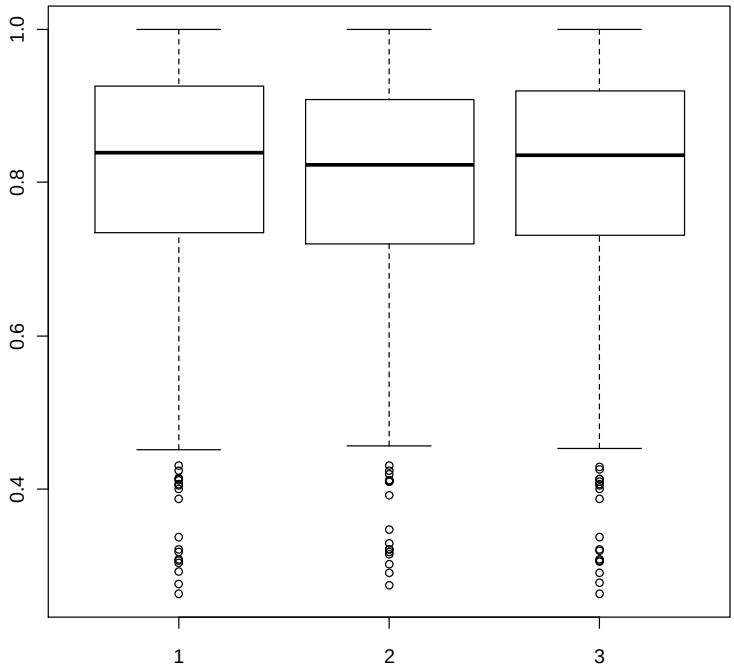

Figura 3.1 (1) $\psi = 0.80, n_x = n_y = 5$, L_p
(2) $\psi = 0.80, n_x = n_y = 5$, L_{mp}*

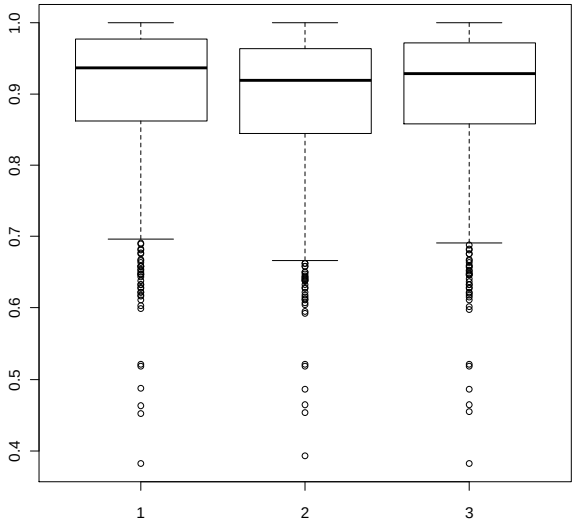

Figura 3.2 (1) $\psi = 0.90, n_x = n_y = 5, L_p$
(2) $\psi = 0.90, n_x = n_y = 5, L_{mp}$*

(3) $\psi = 0.90, n_x = n_y = 5, L_{mp}$

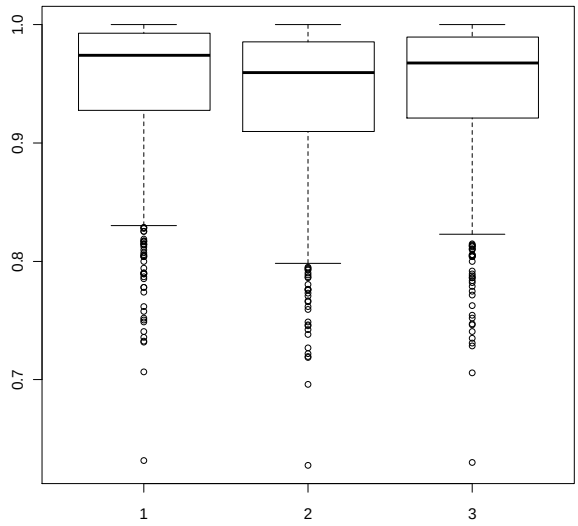

Figura 3.3 (1) $\psi = 0.95, n_x = n_y = 5, L_p$
(2) $\psi = 0.95, n_x = n_y = 5, L_{mp}^*$
(3) $\psi = 0.95, n_x = n_y = 5, L_{mp}$

Per lo studio di simulazione sugli intervalli di confidenza di livello 95% basati su $L_p(\psi)$, $L_{mp}(\psi)$ e $L_{mp}^*(\psi)$, i risultati della simulazione sono riportati nella Tabella 3.2, dove in ogni cella c'è il rapporto tra il numero di volte in cui il valore del test rientra nell'intervallo e il numero di replicazioni.

(n_x, n_y)	Test	$\psi = 0.8$	$\psi = 0.9$	$\psi = 0.95$
(5,5)	$L_p(\psi)$	0.913	0.903	0.918
	$L_{mp}(\psi)$	0.938	0.939	0.953
	$L_{mp}^*(\psi)$	0.946	0.939	0.957
(10,10)	$L_p(\psi)$	0.911	0.911	0.938
	$L_{mp}(\psi)$	0.927	0.931	0.946
	$L_{mp}^*(\psi)$	0.934	0.933	0.951
(20,20)	$L_p(\psi)$	0.944	0.940	0.930
	$L_{mp}(\psi)$	0.949	0.944	0.938
	$L_{mp}^*(\psi)$	0.949	0.946	0.939

Tabella 3.2: tabella del rapporto tra il numero di volte in cui il valore del test rientra nell'intervallo e il numero di ripetizioni

Dalla Tabella 3.2 si nota che la copertura risulta maggiore nelle due verosimiglianze profilo modificate rispetto alla verosimiglianza profilo, per tutte le numerosità campionarie considerate. Tra le due verosimiglianze profilo modificate è migliore quella che presenta il secondo tipo di fattore di aggiustamento (Ventura e Racugno, 2011) anche se non di molto. Quindi possiamo dire che le verosimiglianze profilo modificate risultano una correzione utile della verosimiglianza profilo, in quanto permettono di ottenere risultati che tendono ad eseguire uno studio corretto dei dati disponibili.

BIBLIOGRAFIA

Azzalini A. (2001), *Inferenza statistica, una presentazione basata sul concetto di verosimiglianza,*Springer-Verlag, Berlino.

Azzalini A., Scarpa B. (2004), *Analisi dei dati e data mining*, Springer-Verlag, Berlino.

Barndorff-Nielsen O. E. (1983), *On a formula for the distribution of the maximum likelihood estimator*, Biometrika, 343-365.

Barndorff-Nielsen (1988), *Parametric statistical models and likelihood, Lecture notes in statistics*, Springer, Heidelberg.

Kotz S., Lumelski Y., Pensky M. (2003), *The Stess –Strenght Model and its Generalizations – Theory and Applications*, World Scientific, Singapore.

Severini T. A. (2000), *Likelihood Methods in Statistics*, Oxford University, Oxford.

Pace L., Salvan A. (1996), *Teoria della statistica – Metodi, modelli approssimazioni asintotiche*, CEDAM, Padova.

Pace L., Salvan A. (2001), *Introduzione alla Statistica- II Inferenza, Verosimiglianza, Modelli*, CEDAM, Padova.

Ventura L. e Racugno W. (2011), *Recent advances on Bayesan inference for P(X min Y), Working Papers*, 2010.7, Dipartimento di Scienze Statistiche, Università degli Studi di Padova.